POLYNUCLEOTIDE SEQUENCING METHOD AND KITS THEREFOR

The present invention relates to a method of sequencing a polynucleotide utilising restriction enzymes which cleave the polynucleotide at a site away from the restriction enzyme's recognition site and to kits for use with such a method.

With the advent of genetic engineering it has become possible to isolate polynucleotide fragments and to determine their nucleotide sequence. Typically the polynucleotide fragment of interest is first amplified in order to generate enough sequencing template, prior to determining its polynucleotide sequence. This may be achieved, for example using polymerase chain reaction (PCR) techniques or polynucleotide cloning methodologies.

However, it is generally difficult to sequence large polynucleotide (eg. DNA) fragments (ie. greater than about 500bp-1kbp), due to the limitations of sequencing methodologies. It is often therefore desirable to cleave large fragments into more manageable smaller fragments and to sequence these smaller fragments. The sequences determined can then be reassembled into a single polynucleotide sequence.

One technique of obtaining smaller fragments is known as shotgun cloning. Typically, a large DNA fragment is completely digested, using a frequent cutting restriction enzyme, such as Sau3AI, into much smaller fragments. A vector, for example a plasmid, is digested with a rarer cutting enzyme (e.g. BamHI), so that the vector is cut only

once and so as to give complementary ends to those generated by the frequent cutting enzyme. The small Sau3AI digested polynucleotide fragments are then cloned into the vector to allow sequencing.

However, such a strategy is not attractive because the ends of the DNA fragments produced by digestion are identical and so it is not easy to reassemble them into the order in which they occur in the large fragment without resorting to some form of restriction mapping. Additionally, it is possible to fail to identify colonies containing vectors with very small inserts since such colonies can appear blue using conventional blue/white selection. Unless an accurate restriction map has been determined, it is possible to fail to identify that such small inserts of sequence are missing from the whole sequence and consequently ascertain the polynucleotide sequence of the larger fragment incorrectly.

Thus, it is generally necessary to perform further sequencing experiments in order to confirm the restriction sites and ensure that all fragments have been cloned and sequenced. It will be appreciated that this process can be very time consuming and expensive to perform.

An alternative is to carry out a partial digest, again using a restriction enzyme such as Sau3AI. The partial digest is intended to generate a series of overlapping clones which can be sequenced and the matching sequences aligned so as to form a contiguous overlapping sequence.

However, the conditions for carrying out the partial digestion have to be carefully controlled in order to prevent complete digestion, the control of which can be difficult to achieve. Moreover, a significant amount of overlapping sequence may be generated which may lead to some sections of the DNA being unnecessarily sequenced, which again wastes time and resources.

Another system for sequencing large fragments of DNA is based on the procedure developed by Henikoff (Henikoff, S. (1984) Gene 28, 351), in which exonuclease III (ExoIII) is used to specifically digest DNA from a 5' protruding or blunt-end restriction site. The other end of the DNA is protected from digestion by ExoIII by a 4-base 3' overhang restriction site or by an alpha-phosphorothioate filled end.

Typically ExoIII is added to a sample of linearised vector containing insert DNA and digestion started. Samples of the ExoIII digestion are removed at timed intervals and added to tubes containing SI nuclease, which removes the remaining single-stranded tails. The ends are blunt-ended and ligated to re-circularise the now deletion-containing vectors.

The generation of ordered sets of deletions by this method relies on the uniform digestion rate of ExoIII. However, ExoIII will also digest from nicks in double-stranded DNA. It is therefore important to minimise the proportion of nicked molecules in the starting DNA, by purifying the DNA using special techniques.

Moreover, the ExoIII process is generally only suitable if the restriction enzyme sites which linearise the vector are not present in the insert, the probability of which decreases with increasing insert size. Furthermore the ExoIII process only results in DNA which decreases in size from one end, since the other end is not digested. Thus, subsequent sequencing only generates new sequence from one end.

There is thus the need for a more efficient and easier process which will allow large fragments of polynucleotides (e.g. DNA) to be sequenced.

It is therefore among the objects of the present invention to obviate and/or mitigate at least one of the above described disadvantages.

The present invention provides a method of determining the nucleotide sequence of a polynucleotide, comprising the steps of:

a) cleaving the polynucleotide with a restriction enzyme so as to generate two or more fragments, wherein the restriction enzyme cleaves the polynucleotide at a site away from the restriction enzyme's recognition site so as to generate a cleaved site possessing a recessed 3'-end and a 5'-overhang of undefined sequence;

b) filling-in said recessed 3'-ends so as to form substantially blunt-ended fragments;

c) cloning and sequencing said blunt-ended fragments;

d) pairing matching blunt-ends of said blunt-ended fragments so as to allow said blunt-ended fragments to be

ordered in a contiguous over-lapping arrangement; and

e) reading said nucleotide sequence from said contiguous arrangement.

It is to be understood that the substantially blunt-ended fragments referred to above include fragments with true or perfect blunt-ends (ie blunt-ends which do not possess any overhang) as well as fragments which possess ends with a single-base overhang.

The polynucleotide to be sequenced is generally isolated from the genome with which it is associated and optionally amplified, for example by PCR or cloning into a vector and amplifying the vector in a suitable host. Typically the polynucleotide may be greater than 1kb in length, for example greater than 10kb or greater than 50-100kb in length.

In theory the polynucleotide may be of any length. The suitability of said polynucleotide for sequencing will generally depend on the number and length of restriction fragments which are generated by cleavage with the restriction enzyme.

Although the restriction enzymes cleave double-stranded DNA, the polynucleotide need not initially be double-stranded DNA. The polynucleotide can for example be single-stranded RNA which is converted to double-stranded cDNA by use of reverse transcriptase and DNA polymerase as is well known in the art.

The polynucleotide may be from any desired source. For example, the polynucleotide may be obtained from bacteria, plants, insects, viruses and animals.

The restriction enzymes suitable for use in the present invention specifically generate 5'-overhangs of undefined sequence. The restriction enzyme identifies a constant defined recognition site and cleaves the DNA within an adjacent undefined region which may consist of any sequence. An example of such an enzyme is *Hga*I.

*Hga*I recognises the following recognition site with the recognition sequence shown underlined:

5' G A C G C N N N N N N N N N N N N N 3'

3' C T G C G N N N N N N N N N N N N N 5'

where N represents any nucleotide base (eg. A, C, G or T) and the arrows show the point of cleavage.

Thus, *Hga*I cleavage at this site generates two ends which both possess recessed 3'-ends and 5'-overhangs of undefined sequence, one of which is:

5' G A C G C N N N N N 3'

3' C T G C G N N N N N N N N N N 5'

By convention recognition sequences are often only represented by one strand only, written from 5' - 3'. For enzymes such as *Hga*I, which cleave away from their recognition sequence, the sites of cleavage are indicated by their position, or in parentheses. Thus, the recognition sequence of *Hga*I is often represented as:

GACGC(N)5/10, or

GACGC(5/10)

which means that the enzyme recognises the sequence GACGC and cleaves the DNA within an adjacent region of any sequence, 5 bases away from the end "C" of the recognition sequence on the same strand and 10 bases away on the other strand.

Examples of other restriction enzymes suitable for use in the present invention and their recognition sites are as follows:

*Alw*26I	GTCTC(N)1/5
*Bbv*I	GCAGC(N)8/12
*Bsm*AI	GTCTC(N)1/5
*Bsm*FI	GTCCC(N)10/14
*Bst*71I	GCAGC(N)8/12
*Fok*I	GGATG(N)9/13
*Sfa*NI	GCATC(N)5/9
*Eam*1104I/	
*Ear*I/*Ksp*632I	CTCTTC(N)1/4
*Bbs*I/*Bbv*16II/	
*Bpi*I/*Bpu*AI	GAAGAC(N)2/6
*Bsa*I/*Eco*31I	GGTCTC(N)1/5
*Bsm*B1/*Esp*3I	CGTCTC(N)1/5
*Bsp*MI	ACCTGC(N)4/8
*Gdi*II	CGGCC(A/G)(N)1/5
*Sap*I	GCTCTTC(N)1/4

Any restriction enzyme which generates a 5'-overhang of undefined sequence may be used in the present invention. However, it is preferable that the overhang be 3 or more bases in length in order to minimise the probability of a chance overlap match, as will be explained in detail below.

Typically, the recessed 3'-ends are filled in by employing a DNA polymerase and a mixture of deoxy-nucleotide triphosphates (dNTPs), ie. a mixture containing dATP, dCTP, dGTP and dTTP, so as to generate substantially blunt-ends. DNA polymerases possess the ability to add nucleotides onto an available 3'-OH group of a polynucleotide chain, but cannot add bases to the 5'-phospate group.

The skilled addressee is aware that DNA polymerases that have a "proofreading" function, such as DNA polymerase I, *Pfu* and *Tli* exhibit 3' - 5' exonuclease activity and produce greater than 95% blunt-ended fragments. However, certain thermostable polymerases including *Taq*, *Tfl* and *Tth* polymerase add a single nucleotide, preferentially adenine, to the 3'-end, so as to form a blunt-end possessing a single additional base overhang. However, the single nucleotide overhang can be used to assist with the cloning of the DNA, since perfectly blunt-ended fragments can be more difficult to clone.

The substantially blunt-ended fragments (ie. perfectly blunt-ended fragments or blunt-ended fragments possessing a single base overhang) are cloned into an appropriately digested vector, such as a plasmid, phagemid or phage

cloning vector. Typically the blunt-ended fragments are cloned into a so-called polycloning region of such vectors which possesses a number of unique restriction enzyme sites.

The polycloning region may for example be digested with a restriction enzyme which generates blunt ends, such as *Sma*I or *Hinc*II or alternatively digested with any restriction enzyme which generates a 5'-overhang, since this may also be filled-in by a filling in reaction, to allow cloning of the substantially blunt-ended fragments.

Blunt-ended fragments which possess a single adenine overhang may be cloned into so-called "T-tailed vectors", or "TA cloning vectors" such as the pGEM®-T vector systems available from Promega,Southampton, UK, using techniques previously described in the art (see for example Clark, J. (1988) Nucleic Acids Research 16, 9677 - 9686).

Once the blunt-ended fragments have been cloned their nucleotide sequence may be determined using conventional DNA sequencing methods well known in the art. In particular, the sequence of the previously undefined 5'-overhang region of the cleavage site, which was blunt-ended by the filling-in process, is determined.

Since a single cleavage reaction generates two identically complementary 5'-overhangs, albeit of initially undefined sequence, sequencing of individual clones helps identify which fragment ends were generated by a particular cleavage reaction. This is made possible due to the nature of the restriction enzymes used which generate variable 5'-

overhangs.

The chances of two 5'-overhangs, generated by separate cleavage reactions at different points in the polynucleotide sequence, being accidentally the same, is calculated as 4 (which is the number of possible bases ie A, C, G or T) raised to the power of the length of 5'-overhang. Thus for a restriction enzyme which generates a 3-base 5'-overhang of undefined sequence, the chances of any two separate 5'-overhangs being the same is $1:4^3$ or 1:64. For a restriction enzyme which generates a 5-base 5'-overhang, the chances of any two separate 5'-overhangs being the same is 1:1024. Therefore, providing that relatively few fragments are generated by a particular restriction enzyme, in comparison with the probability of a chance match between any two separate 5'-overhangs, it is possible to pair matching ends with a high degree of certainty that they were generated from the same cleavage reaction at a given point in the polynucleotide sequence.

In this manner it is possible to identify all matching ends by their sequence. The matching ends can then be paired and the fragments ordered so as to allow a contiguous over-lapping arrangement of sequences to be generated, from which the nucleotide sequence of the polynucleotide may be determined. Typically, pairing of the matching ends and ordering of the fragments into a contiguous over-lapping arrangement may be carried out by using a computer program designed for such an application.

Reading of said nucleotide sequence from said contiguous arrangement may then also be carried out by or with the assistance of a computer.

It may be appreciated that the method described herein may be used in conjunction with manual, semi-automated or fully automated sequencing apparatus known in the art.

In manual sequencing the scientist typically reads the sequence off an autoradiograph taken from a gel, on which radioactive or chemiluminescent DNA fragments have been separated according to size by electrophoresis. Such techniques are well known in the art and are described for example in Sambrook, J et al (1989) Molecular Cloning: a laboratory manual, Cold Spring Harbor Laboratory, Cold Spring Harbor, N.Y. The sequence is then conveniently entered into a computer to facilitate observation and/or manipulation of the sequence using appropriate computer software. However, manual sequencing is being circumvented by semi-automated or fully automated sequencing apparatus which can not only determine the sequence of a particular polynucleotide, but can input this information directly into a computer comprising appropriate sequence handling computer software.

It is therefore immediately evident that a computer program designed for pairing of the matching ends and ordering of the fragments into a contiguous over-lapping arrangement may be provided which is suitable for use with the method of the present invention when using manual, semi-automated, and/or fully automated sequencing

apparatus. For example it may be possible to provide suitable software for use in conjunction with a semi-automated or fully automated sequencing apparatus such that the fragments generated using the method of the present invention may be sequenced using a single apparatus linked to a computer comprising the computer software. The sequences of the various fragments are determined using the sequencing apparatus, and the software is able to pair the matching ends and order the fragments into a contiguous over-lapping arrangement. Thereafter the software is able to determine the sequence of said nucleotide from said contiguous arrangement and provided the user with a single nucleotide sequence corresponding to the original polynucleotide.

Thus in a further aspect the present invention provides a computer program for use with the method as described herein, wherein the computer program serves to pair the matching ends of the sequenced fragments and order the fragments into a contiguous overlapping arrangement, thereafter the computer program may read from the contiguous overlapping arrangement and provide the user with the nucleotide sequence of the original polynucleotide. Such a computer program may be provided to a user of the present invention on a computer readable medium such as a floppy disk, CD-ROM or the like. Alternatively semi-automated or fully automated sequence apparatus with a dedicated computer may be provided with the computer program preloaded into the computer's memory.

In order to help better understand the process of pairing the matching ends and ordering the sequences, reference is made to Figure 1 which shows the process schematically.

Part A of Figure 1 shows five fragments (1 to 5) which were generated from a single polynucleotide fragment which had been cleaved with a restriction enzyme as defined above. The fragments have been blunt-ended by filling-in as described, cloned and sequenced. The small regions of sequence corresponding to the 5'-overlaps generated by the restriction enzyme are shown as different symbols. To a high degree of certainty only the ends generated by a particular cleavage reaction will be the same. Thus, for example, the right hand end of fragment 5 matches the left hand end of fragment 3.

By pairing the matching ends of the fragments it is possible to order the fragments in a contiguous overlapping linear arrangement as represented in part B of Figure 1. Once the fragments are ordered as shown in part B, the nucleotide sequence of the original polynucleotide can be easily determined (as shown in part C of Figure 1).

In the example as represented by Figure 1 only two individual ends match with one another. When only a few fragments are generated the likelihood of more than two ends matching is remote. Indeed Table 1 shows the estimated average length of DNA that would be expected before identical restriction sites for each particular restriction enzyme would be observed.

Where there are random overlaps, more than one contiguous arrangement permutation is possible. However, most permutations can be discounted immediately, for example, permutations that produce a circular contiguous arrangement for a DNA fragment that is linear. Additionally the polynucleotide could be cleaved using a different restriction enzyme or a partial digest performed in order to assist in ordering the fragments.

The present process may also be used to conduct restriction mapping of a polynucleotide. To achieve this, it is not necessary to sequence the entire length of each fragment, only the blunt-ends generated from the restriction enzyme digestion and filling-in reaction need be sequenced. It is then possible to order the fragments as described above in order to generate a restriction map.

In another aspect the present invention provides a kit suitable for use in any of said processes according to the present invention, the kit comprising at least one restriction enzyme as defined herein together with a DNA polymerase or polymerases for the filling-in and/or sequencing reactions. Other components such as dNTPs, a T-tailed vector, competent cells, sequencing reagents and the like may also be included as appropriate. In addition a computer program in a machine readable form such as a computer disk or CD-ROM may be provided for pairing the matching ends and ordering the fragments into a contiguous overlapping arrangement and thereafter providing the nucleotide sequence of the polynucleotide.

The present invention will be further described and understood with reference to the following non-limiting Examples section.

Examples Section

Materials & Methods

1. Restriction Enzyme Digests

All restriction enzyme digests were performed on pure DNA using restriction enzymes supplied by Promega (Promega, Southampton, UK) or New England Biolabs (New England Biolabs, Hitchin, UK). Incubation conditions were 37°C for a minimum of 1 hour using the appropriate buffer supplied by Promega or New England Biolabs. Following digestion DNA was run on an agarose gel and gel extracted using QiaexII gel extraction kit (Qiagen, Crawley, UK).

2. Extraction of DNA from agarose gels with the QIAEX II gel extraction kit

All DNA extracted from gels was purified using the QIAEX II DNA gel extraction kit according to the manufacturer's instructions. Briefly, three volumes of 'QX-1' buffer and 10µl of QiaexII DNA binding beads were added to each gel plug. The plugs were dissolved by warming to 50°C during which time the beads were kept suspended by vortexing every 2 min. After 10 min the beads were pelleted by a 20s centrifugation in a benchtop centrifuge. The supernatant was removed and the pellet washed in 500µl 'QX-1' buffer, resuspended, and then

pelleted in the same manner as above. The pellet was then washed, resuspended and pelleted similarly in an ethanolic wash 'PE' buffer. The pellet was then allowed to dry for 10 min and then eluted in 20µl of water. This DNA was typically contaminated with ethanol and so was subsequently purified by ethanol precipitation.

3. Ethanol precipitation

To the volume of DNA to be ethanol precipitated, 0.1 volume 3M sodium acetate was added and 2 volumes of 100% ethanol. The vial was mixed and incubated at -80°C for 30 minutes. The precipitated suspension was centrifuged at 11000rpm in a Jouan (MR1812) refrigerated centrifuge for 10 min to pellet the DNA. The supernatant was aspirated and 1ml of 70% ethanol added. The DNA was pelleted again by centrifugation at 11000rpm in the refrigerated centrifuge for 5 min, the supernatant aspirated and the pellet allowed to air-dry for 5-10 minutes. The DNA was resuspended in TE buffer and the purity of the DNA checked by UV absorption at 260nm and 280nm, where A260/A280=1.8 for pure plasmid DNA.

4. Generation of plasmid DNA

Plasmid DNA was prepared using maxiprep and miniprep kits (Promega, Southampton, UK). A brief protocol for a Promega maxiprep kit is given below.

Promega maxi-prep

A culture was set up by stabbing a toothpick into frozen glycerol stocks and adding it to 400ml of ampicillin (50μg/ml) LB medium. The culture was incubated overnight at 37°C in a rotating incubator at 200rpm.

Preparation of cleared lysate

The culture was then poured equally into 250ml Beckman centrifuge tubes and pelleted at 9500g for 10 mins at room temperature in a JA-14 rotor. Each pellet was resuspended in 7.5ml 'Resuspension solution' using a heat-sealed 5ml pipette to manually disrupt the pellet. These suspensions were combined. To the combined 15ml, 15ml 'Cell Lysis' solution was added and mixed by inversion. Lysis was allowed to complete (up to 20 min) and then 15ml of 'Neutralisation solution' was added and immediately mixed by inversion.

The suspension was centrifuged at 14,000g for 15 min at room temperature. The cleared supernatant was transferred to a new container.

Plasmid DNA precipitation

0.6 volumes of isopropanol was now added and mixed by inverting several times. The DNA was pelleted by centrifugation at 14,000g for 15 mins at room temperature. The supernatant was discarded and the DNA pellet resuspended in 2ml TE.

Plasmid purification

One Maxicolumn was inserted into a vacuum manifold. 10ml of well-shaken pre-warmed 'DNA purification resin' was added to the DNA/TE solution and then this slurry was added to the maxicolumn. A vacuum was applied to draw the slurry through. The DNA/resin contained was rinsed with 13ml of 'Column wash solution' and immediately added to the column under vacuum. A final wash of 12ml of 'Column wash solution' was then added to the column. The resin was rinsed with 5ml of 80% isopropanol under vacuum.

The resin was dried by centrifuging the column in its 50ml conical tube in a bench-top clinical centrifuge at 2,500 rpm (1300 g) for 5 min. It was then transferred to a new 50ml conical centrifuge tube. 1.5 ml pre-heated water (65-70°C) was applied to the tube. After 1 minute this water was centrifuged out of the column using the conditions above.

DNA solution was stored -20°C.

5. Cloning DNA

Phosphatase treatment of DNA

If appropriate (ie not necessary for TA-cloning vectors) prior to ligation of an insert into a vector, the plasmid DNA was treated with calf intestinal alkaline phosphatase (CIAP) if the vector had been digested with a single restriction enzyme. The CIAP removes the 5' phosphate groups and thus prevents recircularization of the vector during ligation.

Reaction mix

The following was added to a microcentrifuge tube:

vector DNA

CIAP 10x buffer

CIAP

dH_2O

This was mixed gently and incubated for 1 hour at 37°C. CIAP was removed prior to ligation by phenol/chloroform extraction.

Double stranded DNA ligation

Double stranded DNA with cohesive ends was ligated into 100ng vector by adding 1 unit of T4 DNA ligase (Promega) to 1:1 and 1:3 ratios of vector and insert DNA in 19.5µl 1 x Ligase buffer (10X T4 DNA ligase buffer is 30mM Tris-Hcl, pH7.8. 100mM $MgCl_2$, 100mM DTT, 10mM ATP). This reaction was incubated at 14°C overnight. The ligase buffer was aliquoted to prevent degradation of ATP.

6. TA Cloning

Sample preparation

The DNA precipitate from an ethanol precipitation was resuspended in a volume such that the ratio of concentration of the average sized insert to vector would be 3:1 in the ligation.

Ligation

TA cloning depends on a property which certain polymerases possess in transferring single adenines onto the 3' end of blunt-ended DNAs. Vectors carrying complementary T overhangs can ligate with these DNAs very efficiently because neither molecule can circularise thus promoting intermolecular reactions. TA cloning was performed using the Original TA cloning® kit or the Eukaryotic TA cloning® kit (both available from Invitrogen BV, NV Leek, Holland) (Bidirectional) as required. The ligation reaction is carried out essentially as above, using the supplied precut vector containing the T overhang.

TA cloning: transformation

An aliquot of frozen competent cells (either invαF'or TOP10F' supplied by Invitrogen) was thawed slowly on ice. 2μl of ligation reaction and 2μl of 0.5Mβ-Mercaptoethanol was added to the tube, mixed with the pipette tip and incubated on ice for 30 min. The cells were then heat shocked for 30s at 42°C and incubated on ice for a further 2 min. 250μl of SOC broth was then added and the transformed cells incubated at 37°C for 60 min with shaking (225rpm). 100μl of the culture was plated on a 10cm agar plate containing 50μg/ml ampicillin. Transformed colonies were identifiable in the 'Original' TA cloning vector (pCR2.1) using blue/white colour selection because of insertion into the β-Galactosidase gene. Colour selection was not possible in the eukaryotic TA cloning expression

vector (pCR3.1). In either case white colonies were picked, PCR screened to ensure an insert was present and glycerol stocks made of positive colonies.

7. DNA sequencing with the ABI sequencer

Protocol for cycle sequencing

Samples for sequencing taken from maxi-preps or mini-preps were mixed with the TaqDyeDeoxy Terminator (Applied Biosystems, Foster City, CA, USA) reaction premix.

Reaction mix:

Reaction premix (contains buffer, polymerase, dNTPs, ddNTPs, magnesium)	8µl
ds DNA template	400ng
Primer (for ds DNA)	3.2pmol
H_2O	to 20µl

Sequencing reactions so prepared were subjected to thermal cycling using the following conditions:

Cycles (25)	Denaturation	96°C for 30s
	Annealing*	47°C for 15s
	Extension	60°C for 4min

* This segment temperature was variable according to the primer used. The temperature shown was that used for the T7 sequencing primer (taatacgactcactataggg) and the pCR2.1

upstream primer (agctatgaccatgattacg).

Reaction products were concentrated by ethanol precipitation and the pellets sent to the Glasgow University Molecular Biology Support Unit for gel electrophoresis and sequence determination using an Applied Biosystems fluorescent sequencer model 393A.

Sequence analysis

Routine DNA sequence handling and analysis was performed on the Gene Jockey II program (Biosoft, Cambridge, UK)

Example 1

Generation of *HgaI* digested fragments

A 2.4kb Xho II fragment which had been cloned into a plasmid vector was re-excised using flanking EcoR I sites and the resulting fragment was digested, as described above, with an 1 unit excess amounts appropriate or of *HgaI* restriction endonuclease.

Digestion generated four separate fragments, two of 0.4kb, one of 0.7kb and one of 0.9kb. The fragments where separated by gel electrophoresis and purified from appropriate gel fragments, as described above.

Example 2

Cloning the *HgaI* digested fragments

The recessed 3'-ends generated by *HgaI* and Xho II digestion of the purified fragments were filled in using *Taq* polymerase and dNTPs as outlined below.

Filing in reaction as follows:

Gel purified DNA (dissolved in water)	to final volume 50µl
Taq polymerase 10X reaction buffer (Promega)	5µl
dNTPs(2mM of each dNTP)	5µl
25mM $MgCl_2$	3µl
Taq polymerase (Promega)	2.5 units

Incubated at 65°C for 10 min.

A single TA ligation reaction was set up with the filled-in fragments and "original" TA vector as described above.

After ligation was completed invαF' competent cells were transformed and plated on agar plates, containing IPTG/Xgal for blue/white colour selection, and colonies allowed to develop.

Plasmid DNA was prepared from a selection of colonies and screened using PCR techniques in order to ensure insert was present in the vector.

Insert containing plasmid DNA was then prepared for DNA sequencing.

Example 3

Sequencing the cloned fragments and ordering the sequences into a contiguous arrangement

Sequencing of the plasmid DNA was carried out as described above and the sequence obtained subjected to sequence analysis.

Figure 2 shows the short unique overlaps which were introduced into the polynucleotide fragment following digestion with *Hga*I. The solid underlines show the *Hga*I recognition sequences and the bold GATC motif at the ends of fragments 1 and 4 are the ends of the *Xho*II fragment following TA cloning (recognition sequence RGATCY where R = G/A and Y = C/T).

The entire sequence of the fragments is not shown since it is not necessary for the understanding of the underlying principle of the present invention. These short unique overlaps allowed the two 0.4kb, one 0.7kb and one 0.9kb fragments to be ordered into a contiguous overlapping arrangement as shown in Figure 3a. Figure 3b shows the *Hga*I restriction map of the original polynucleotide fragment.

As can be seen from Figure 2 and Figure 3b, the junction between fragments 3 and 4 was more complex than the other junctions, due to two *Hga*I restriction sites being extremely close to each other. However, in such a situation, digestion at one site effectively destroys the other such that an overlap can still be discerned from some clones.

TABLE 1

Restriction	Recognition	Frequency of cutting (bp)	Probability of chance overlap match[1] (bp)	Length of DNA before match[2] (kb)
Alw26I BbvI BsmAI BsmFI Bst71I FokI[3] SfaNI	GTCTC(N)1/5 GCAGC(N)8/12 GTCTC(N)1/5 GTCCC(N)10/14 GCAGC(N)8/12 GGATG(N)9/13 GCATC(N)5/9	1 in 512	1 in 256	131Kb
HgaI	GACGC(N)5/10	1 in 512	1 in 1024	524Kb
Eam1104I/ EarI/Ksp632I	CTCTTC(N)1/4	1 in 2048	1 in 64	131Kb
BbsI/Bbv16II/ BpiI/BpuAI BsaI/Eco31I BsmBI/Esp3I BspMI GdiII[4]	GAAGAC(N)2/6 GGTCTC(N)1/5 CGTCTC(N)1/5 ACCTGC(N)4/8 CGGCCR(N)1/5	1 in 2048	1 in 256	524Kb
SapI	GCTCTTC(N)1/4	1 in 8192	1 in 64	524Kb

[1] Probability of chance match between two unrelated overhangs

[2] Estimated length of DNA before a random match between unrelated recognition sequences. (Example calculation: Alu26I will cut on average once every 512bp. Each overhang has a 1 in 256 chance of matching. Thus estimated the length of DNA before two identical recognition sequences are observed is:
the frequency of cutting x the chances of matching
ie. 512bp x 256bp = 131Kb.